Quest for Structured Representation for 3D Reconstruction and Generation

Shenghua Gao

University of Hong Kong

MACHINE PERCEPTION OF THREE-DIMENSIONAL SOLIDS

by

LAWRENCE GILMAN ROBERTS

S.B., Massachusetts Institute of Technology
(1961)

M.S., Massachusetts Institute of Technology
(1961)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1963

Larry Roberts (December 21, 1937 – December 26, 2018) was an American computer scientist and Internet pioneer.

As a program manager and later office director at the Advanced Research Projects Agency, Roberts and his team created the ARPANET using packet switching techniques invented by British computer scientist Donald Davies and American engineer Paul Baran.[4][5] The ARPANET's principal designer was Bob Kahn who worked at Bolt Beranek and Newman (BBN). Roberts asked Leonard Kleinrock to apply mathematical methods to model and measure the performance of the network. Subsequent ARPA research on communication protocols for internetworking led to the development of the modern Internet.

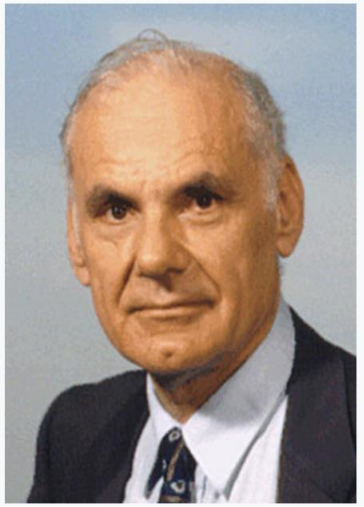
Roberts later was CEO of the commercial packet-switching network Telenet, the first public data network in North America.

Early life and education [edit]

Lawrence Gilman Roberts, who was known as Larry, was born and raised in Westport, Connecticut.[6] He was the son of Elizabeth (Gilman) and Elliott John Roberts, both of whom had doctorates in chemistry. It is said that during his youth, he built a Tesla coil, assembled a television, and designed a telephone network built from transistors for his parents' Girl Scout camp.[7]

Roberts attended the Massachusetts Institute of Technology (MIT), where he received his bachelor's degree (1959), master's degree (1960), and Doctor of Philosophy (Ph.D., 1963),[8] all in electrical engineering.[7] Due to his Ph.D. thesis "Machine Perception of Three-Dimensional Solids"[8] he is known as the father of computer vision.[9][10]

Lawrence Roberts



Roberts in 2017

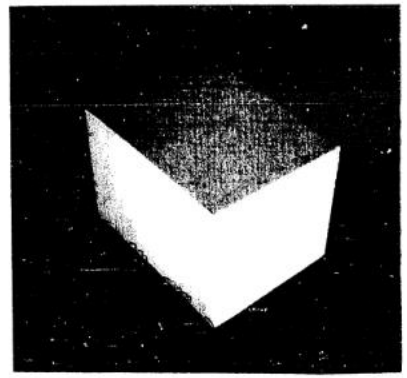
Born	Lawrence Gilman Roberts December 21, 1937 Westport, Connecticut, U.S.
Died	December 26, 2018 (aged 81) Redwood City, California
Alma mater	Massachusetts Institute of Technology
Known for	ARPANET, founding father of the Internet
Awards	Internet Hall of Fame, 2012 · IEEE Computer Pioneer Award · IEEE Computer Society W. Wallace McDowell Award · ACM SIGCOMM Award · Harry H. Goode Memorial Award · International Engineering

MACHINE PERCEPTION OF THREE-DIMENSIONAL SOLIDS

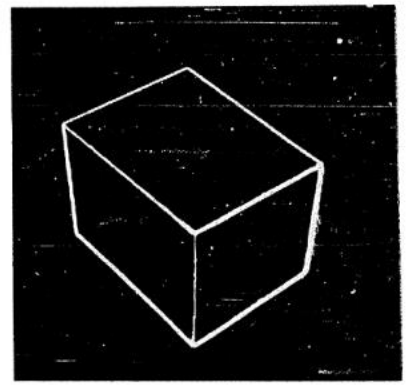
ABSTRACT

In order to make it possible for a computer to construct and display a three-dimensional array of solid objects from a single two-dimensional photograph, the rules and assumptions of depth perception have been carefully analyzed and mechanized. It is assumed that a photograph is a perspective projection of a set of objects which can be constructed from transformations of known three-dimensional models, and that the objects are supported by other visible objects or by a ground plane. These assumptions enable a computer to obtain a reasonable, three-dimensional description from the edge information in a photograph by means of a topological, mathematical process.

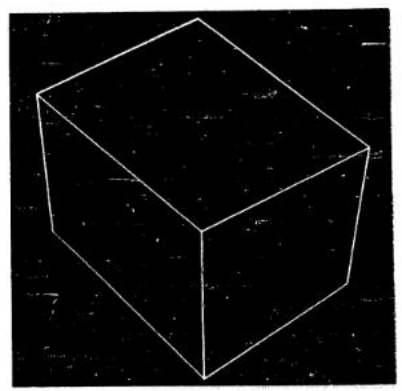
A computer program has been written which can process a photograph into a line drawing, transform the line drawing into a three-dimensional representation, and finally, display the three-dimensional structure with all the hidden lines removed, from any point of view. The 2-D to 3-D construction and 3-D to 2-D display processes are sufficiently general to handle most collections of planar-surfaced objects and provide a valuable starting point for future investigation of computer-aided three-dimensional systems.



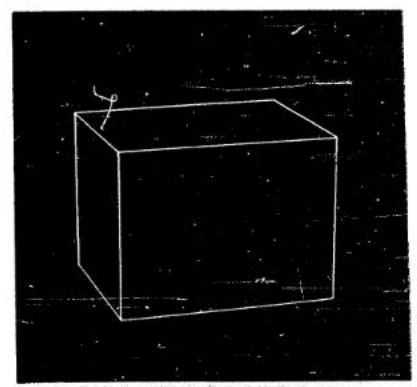
A. Original Picture



B. Differentiated Picture



C. Line Drawing

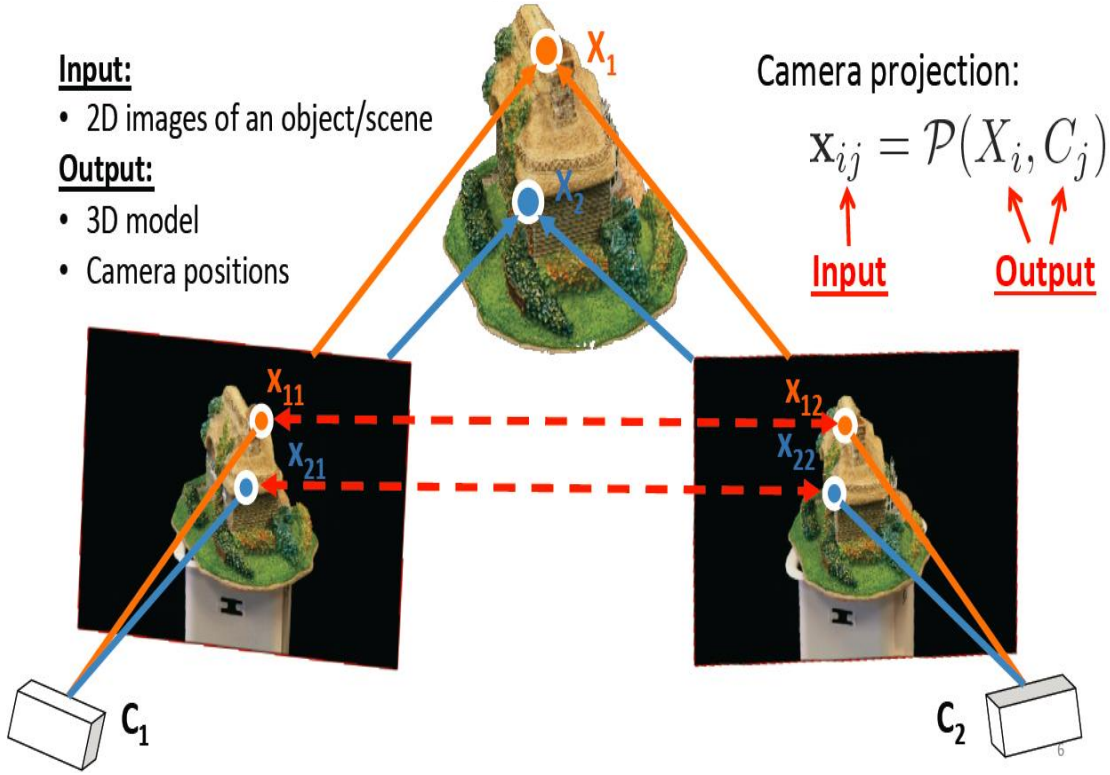


D. Rotated View

Pictures 2A - 2D: Single Model: Reduction of photograph to line drawing and display of 3-D construction from another viewpoint.

Pioneering works on 3D Reconstruction

1. **Ullman (1979)**: Shimon Ullman's paper, *"The Interpretation of Structure from Motion"* (MIT AI Memo 476, 1979), is a seminal work that formalized the computational principles of inferring 3D structure from motion cues. This laid the groundwork for structure-from-motion (SFM) techniques.
2. **Marr and Poggio (1979)**: David Marr and Tomaso Poggio's *"A Computational Theory of Human Stereo Vision"* (Proceedings of the Royal Society of London, 1979) introduced a cooperative algorithm for stereo disparity, foundational for stereo-based 3D reconstruction.
3. **Horn (1977)**: Berthold K.P. Horn's work on *"Understanding Image Intensities"* (Artificial Intelligence, 1977) and subsequent papers explored shape-from-shading, an early method for inferring 3D shape from 2D intensity variations.
4. **Longuet-Higgins (1981)**: H. Christopher Longuet-Higgins' *"A Computer Algorithm for Reconstructing a Scene from Two Projections"* (Nature, 1981) introduced the "eight-point algorithm" for reconstructing scenes from two views, a cornerstone of multi-view geometry.



Point-matching based 3D reconstruction

- 3D reconstruction is an optimization problem!
- Objective function of reprojection

$$\min \sum_{ij} ||X_{ij} - P(\mathrm{X}_i, C_j)||$$

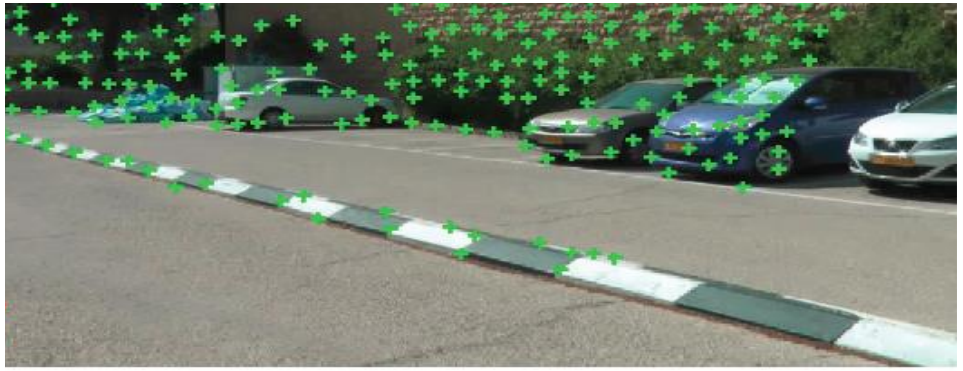
- Current methods can well handle…
  - Millions of images
  - Varying camera poses
  - Different lighting conditions
  - Image noises

San Marco Square, 14,079 images, 4,515,157 points

Building Rome in a Day. Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz and Richard Szeliski. Communications of the ACM, 2011.

But in practice, feature point based matching approaches still fail

Repetitive and symmetric patterns



Textureless areas

Dynamic scenes

The role of structure in human 3D perception

- Structure: **spatial relationships** among multiple points, lines, patches, etc.
- Human perceives 3D space by recognizing **many types of structure** in the scene: *parallelism, planar surfaces, regular shapes, repetitive patterns, symmetry, self-symmetry,…*

Geometric Reasoning for Single Image Structure Recovery.
David C. Lee, Martial Hebert, and Takeo Kanade. CVPR 2009.

Structure learning for 3D reconstruction

Structure learning

- Line detection
- Plane detection
- Room layout estimation

global structure vs. local structure

Challenges in feature matching

- Feature mismatching
- Textureless regions
- Dynamic objects

Line detection

Plane detection

Room layout estimation

Learning to Parse Wireframes in Images of Man-Made Environments

CVPR 2016

Kun Huang[1], Yifan Wang[1], Zihan Zhou[2], Tianjiao Ding[1], Shenghua Gao[1], and Yi Ma[3]
[1]ShanghaiTech University {huangkun, wangyf, dingtj, gaoshh}@shanghaitech.edu.cn
[2]The Pennsylvania State University zzhou@ist.psu.edu
[3]University of California, Berkeley yima@eecs.berkeley.edu

Abstract

In this paper, we propose a learning-based approach to the task of automatically extracting a "wireframe" representation for images of cluttered man-made environments. The wireframe (see Fig. 1) contains all salient straight lines and their junctions of the scene that encode efficiently and accurately large-scale geometry and object shapes. To this end, we have built a very large new dataset of over 5,000 images with wireframes thoroughly labelled by humans. We have proposed two convolutional neural networks that are suitable for extracting junctions and lines with large spatial support, respectively. The networks trained on our dataset have achieved significantly better performance than state-of-the-art methods for junction detection and line segment detection, respectively. We have conducted extensive ex-

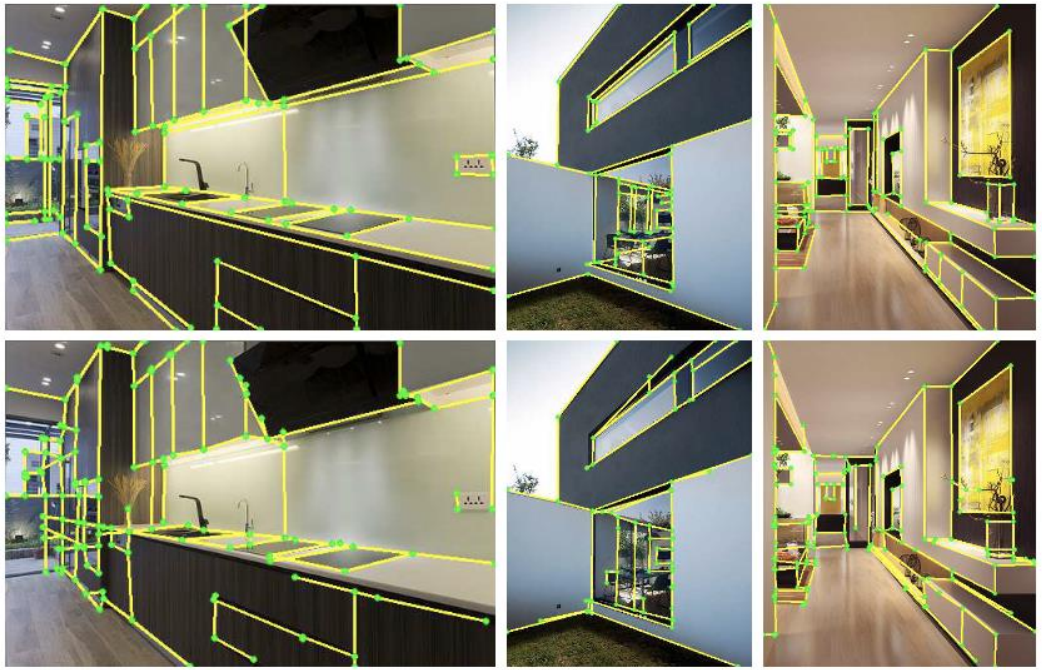
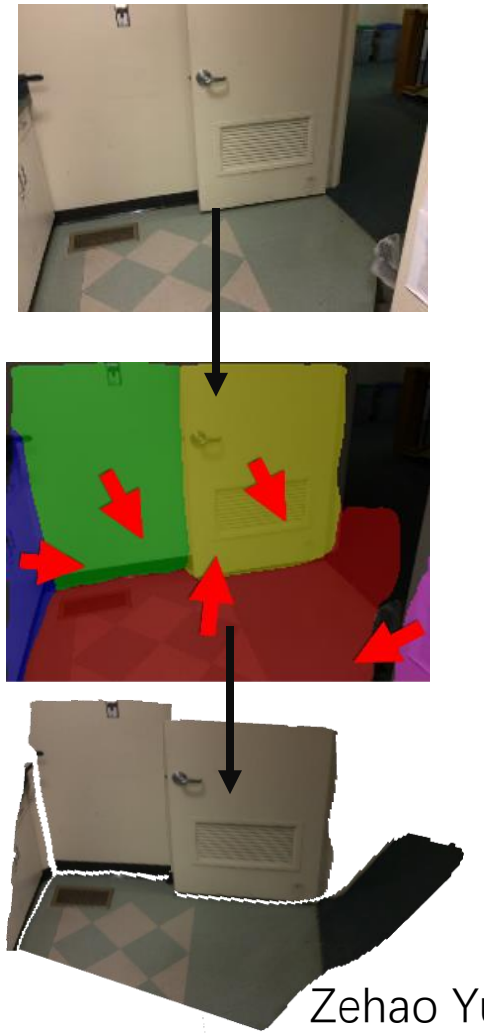


Figure 1. **First row:** Examples of typical indoor or outdoor scenes with geometrically meaningful wireframes labelled by humans; **Second row:** Wireframes automatically extracted by our method.

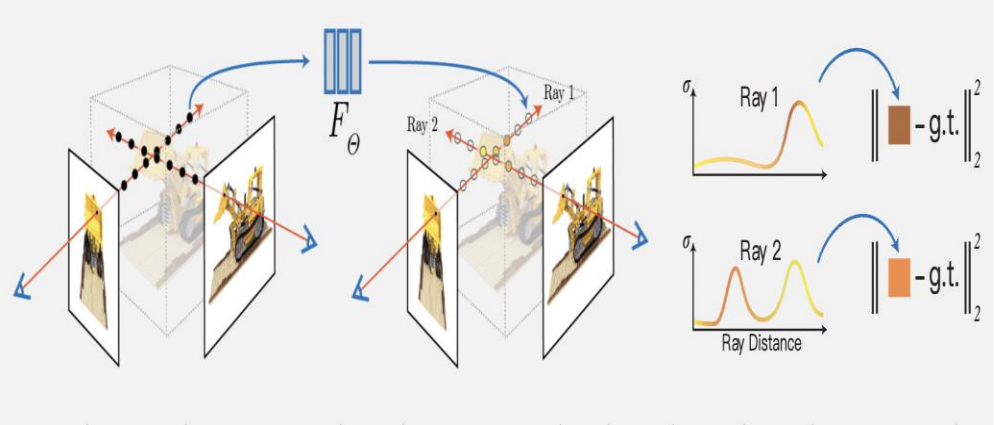
Piecewise planar based 3D representation

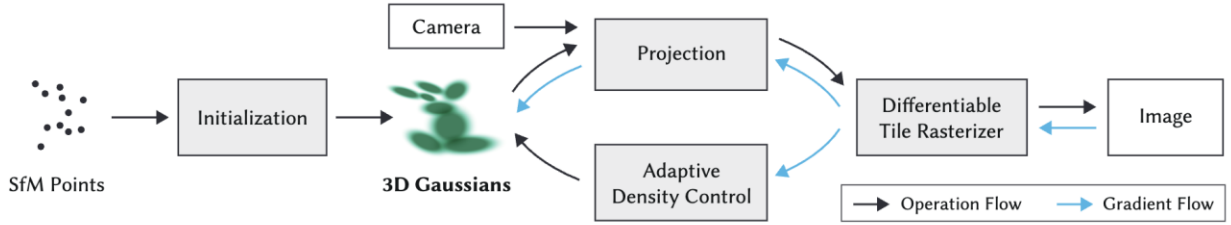
3D plane = 2D segmentation map + 3D parameters



Zehao Yu et al, Single-Image Piece-wise Planar 3D Reconstruction via Associative Embedding CVPR 2016

Neural Radiance Field (NeRF) and 3DGS





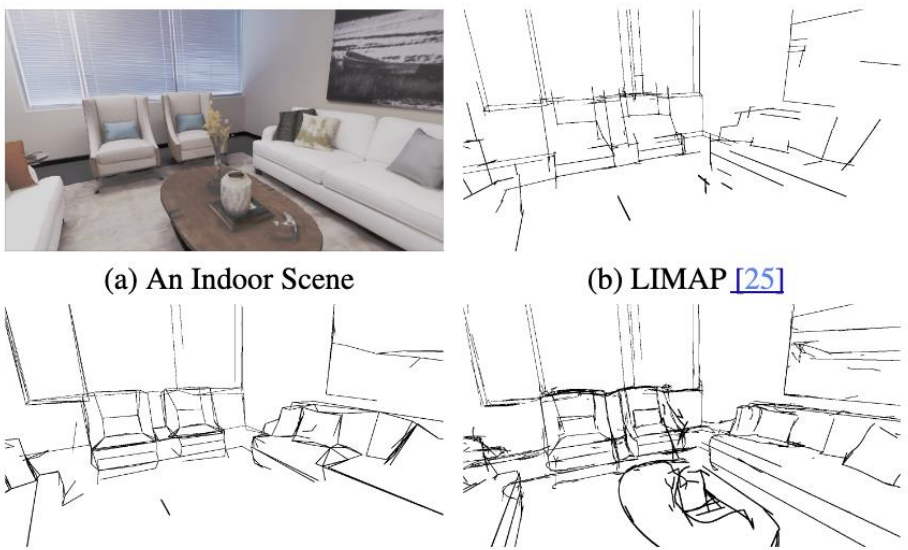
Ben Mildenhall et al, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020
Kerbl et al, 3D Gaussian Splatting for Real-Time Radiance Field Rendering, Siggraph 2023
Perspective Accurate Splatting. Zwicker et al. 2004

3D Neural Edge Reconstruction

Lei Li[1] Songyou Peng[1,2†] Zehao Yu[3,4] Shaohui Liu[1] Rémi Pautrat[1,6]
Xiaochuan Yin[5] Marc Pollefeys[1,6]
[1]ETH Zurich [2]MPI for Intelligent Systems, Tübingen [3]University of Tübingen
[4]Tübingen AI Center [5]Utopilot [6]Microsoft
neural-edge-map.github.io

Abstract

Real-world objects and environments are predominantly composed of edge features, including straight lines and curves. Such edges are crucial elements for various applications, such as CAD modeling, surface meshing, lane mapping, etc. However, existing traditional methods only prioritize lines over curves for simplicity in geometric modeling. To this end, we introduce EMAP, a new method for learning 3D edge representations with a focus on both lines and curves. Our method implicitly encodes 3D edge distance and direction in Unsigned Distance Functions (UDF) from multi-view edge maps. On top of this neural representation,



(a) An Indoor Scene (b) LIMAP [25]

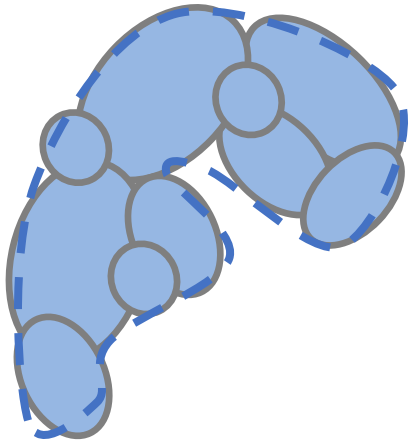
(c) NEAT [64] (d) EMAP (**Ours**)

3D Gaussian Splatting

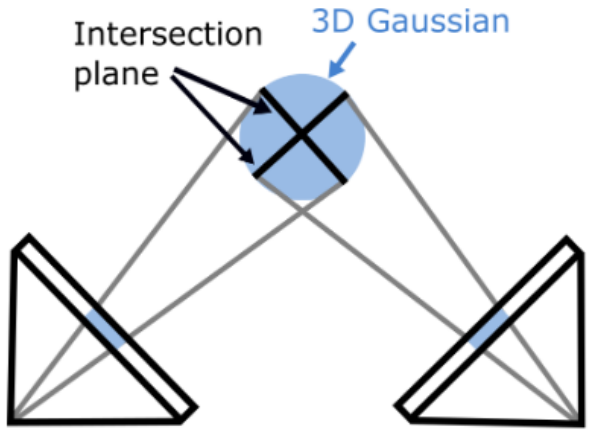
Challenges in Surface Reconstruction

- **Representation:** Volumetric Representation
- **Rendering:** Inaccurate Projection

Reference



Volumetric Density

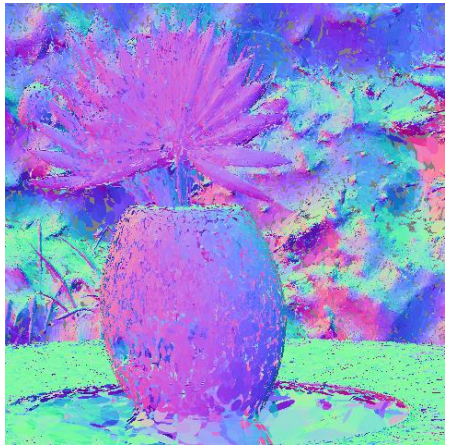


Varying sliced 2D Gaussian

Geometry Reconstruction

2D Gaussian Splatting For Geometrically accurate Radiance Fields

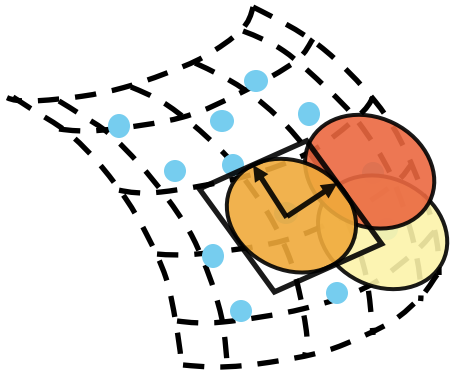




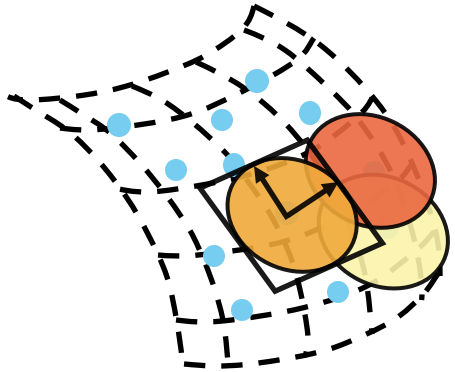
Surfels Splats Mesh (TSDF fusion)

Huang Binbin, et al, 2D Gaussian Splatting For Geometrically accurate Radiance Fields, SIGGRAPH 2024
Pfister et al, Surfels: Surface Elements as Rendering Primitives, SIGGRAPH 2000

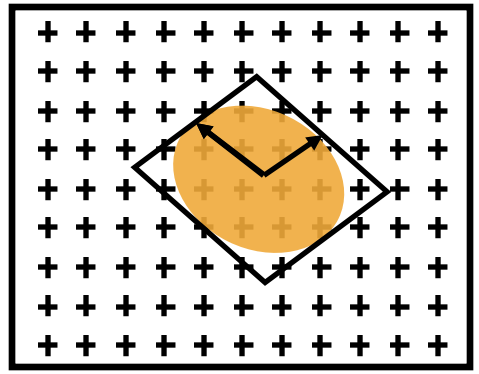
2DGS



1. Consider **oriented points (surfels)** as discrete samples of a texture function on a surface.



2. A 2D Gaussian kernel is defined on **object space** to recover continuous signal.



3. We reconstruct the surface in **screen space.**

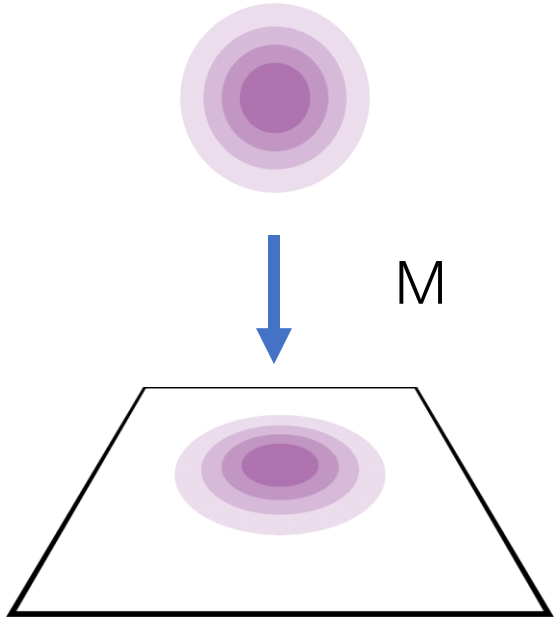
Projection

Surface splatting – Zwicker et al. 2001
(EWA – Elliptical Weighted Average)

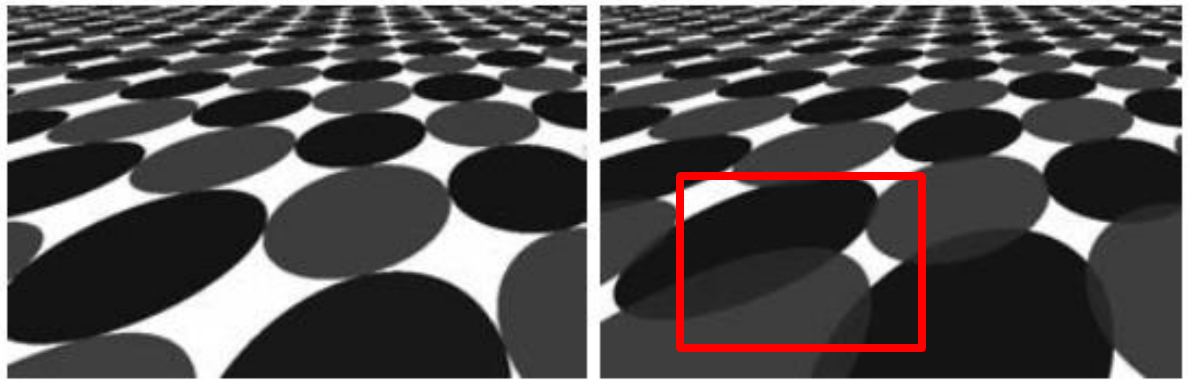
How to render 2D Gaussian?

- **Projection-based:** Perspective Distortion，Inaccurate depth

EWA Surface Splatting



Ground Truth **EWA Splatting**



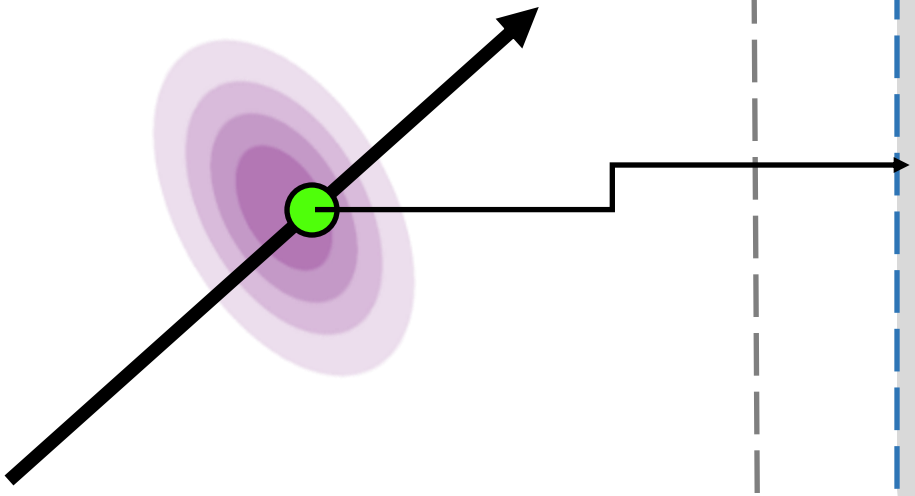
Perspective Distortion

Perspective Accurate Splatting. Zwicker et al. 2004

How to render 2D Gaussian?

- Projection-based: Perspective Distortion， Inaccurate depth
- **Ray-Splat intersection: Perspective correct, Depth correct**

Ray-Splat Intersection



- Evaluate Gaussian on the tangent uv-plane

1. Parameterize a ray with planes $(\boldsymbol{h_x}, \boldsymbol{h_y})$
2. Transform a plane through inverse transpose $(M^{-1})^{-T} = M^T$, eliminate unstable inverse:

$$\mathbf{h}_u = (\mathbf{WH})^{\mathrm{T}}\mathbf{h}_x \quad \mathbf{h}_v = (\mathbf{WH})^{\mathrm{T}}\mathbf{h}_y$$

3. Find the intersection

$$u(\mathbf{x}) = \frac{\mathbf{h}_u^2\mathbf{h}_v^4 - \mathbf{h}_u^4\mathbf{h}_v^2}{\mathbf{h}_u^1\mathbf{h}_v^2 - \mathbf{h}_u^2\mathbf{h}_v^1} \qquad v(\mathbf{x}) = \frac{\mathbf{h}_u^4\mathbf{h}_v^1 - \mathbf{h}_u^1\mathbf{h}_v^4}{\mathbf{h}_u^1\mathbf{h}_v^2 - \mathbf{h}_u^2\mathbf{h}_v^1}$$

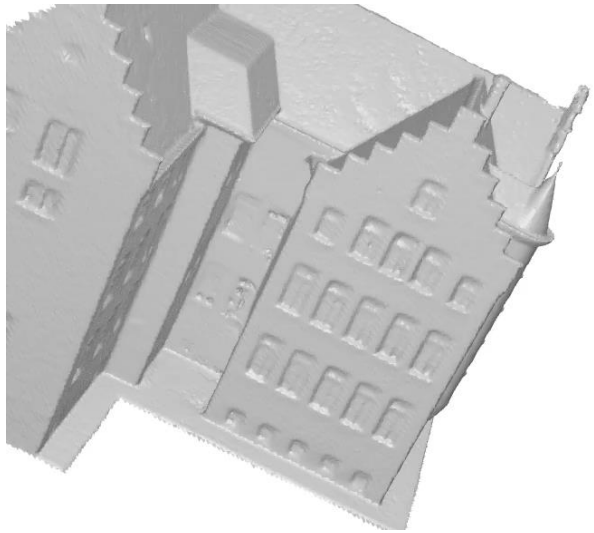
A hardware architecture for surface splatting. Weyrich et.al 2006

Performance Comparison - Geometry

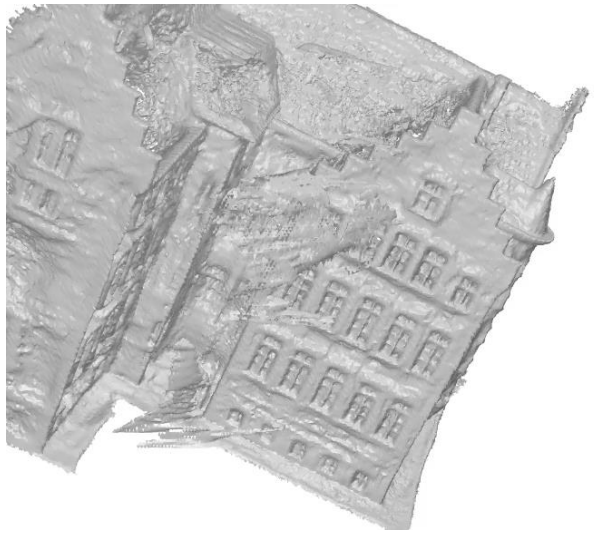
1. High-quality appearance and geometry.

Scan 24

Reference



Ours



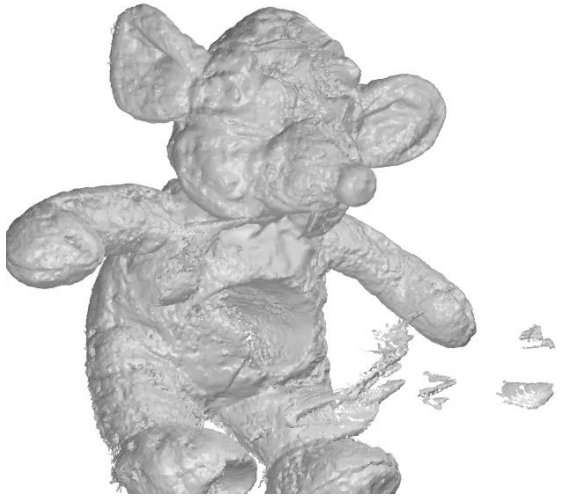
3DGS

Performance Comparison - Geometry

1. High-quality appearance and geometry.

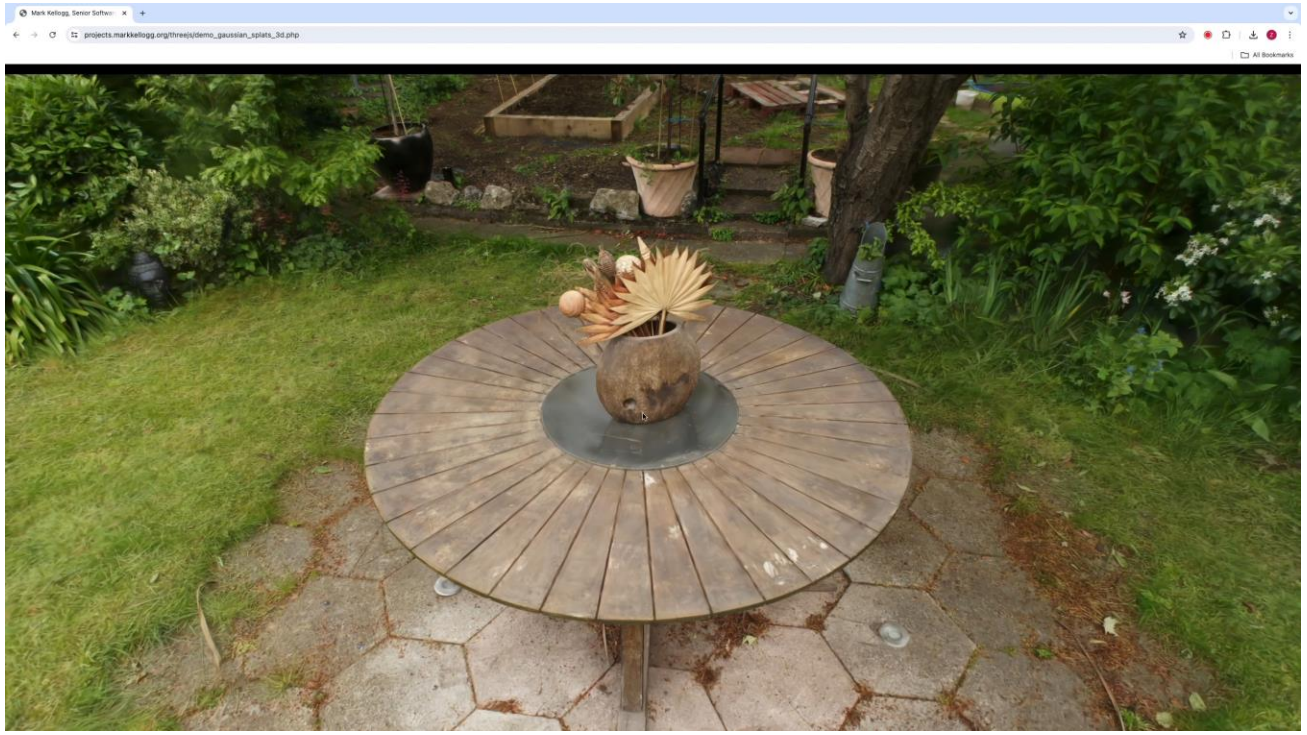
Reference

Ours

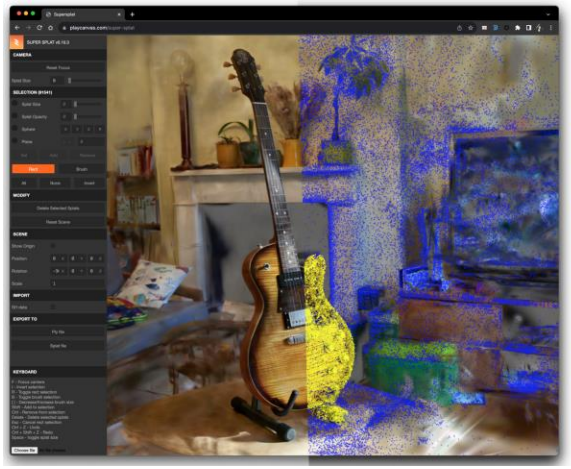


3DGS

Community Support



Web Viewer



Online Editor

gsplat

Efficient Backend

https://projects.markkellogg.org/threejs/demo_gaussian_splats_3d.php (Splat Viewer)
https://github.com/playcanvas/supersplat (Splat Editor)
https://github.com/nerfstudio-project/gsplat (Efficient Backend)

Applications

4D Reconstruction

ViDu4D [Wang et al. 2024]

Video Editing

3D Street Unveiler [Xu et al. 2024]

Dynamic Inverse Rendering

GS-IA [Zhao et al. 2024]

Generalizable Reconstruction

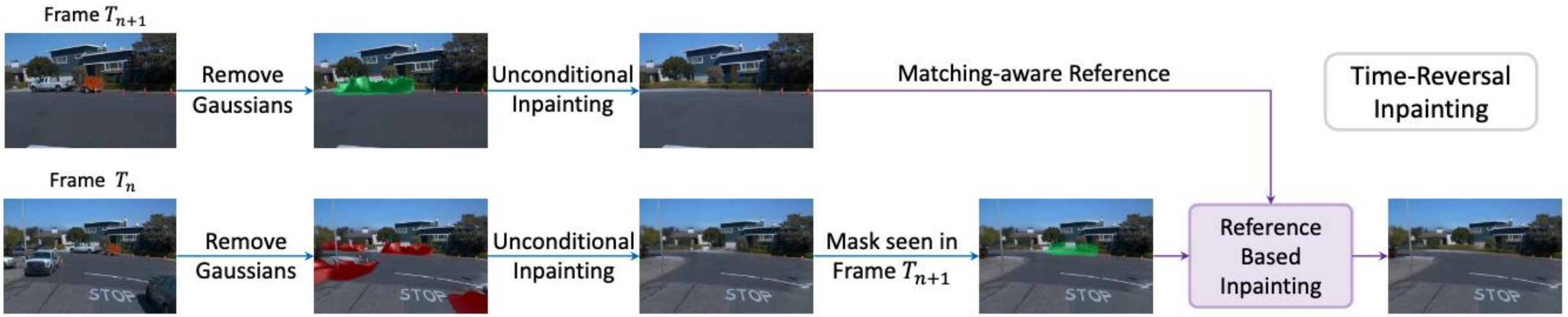
LaRa [Chen et al. 2024]

StreetUnveiler (ICLR 2025)

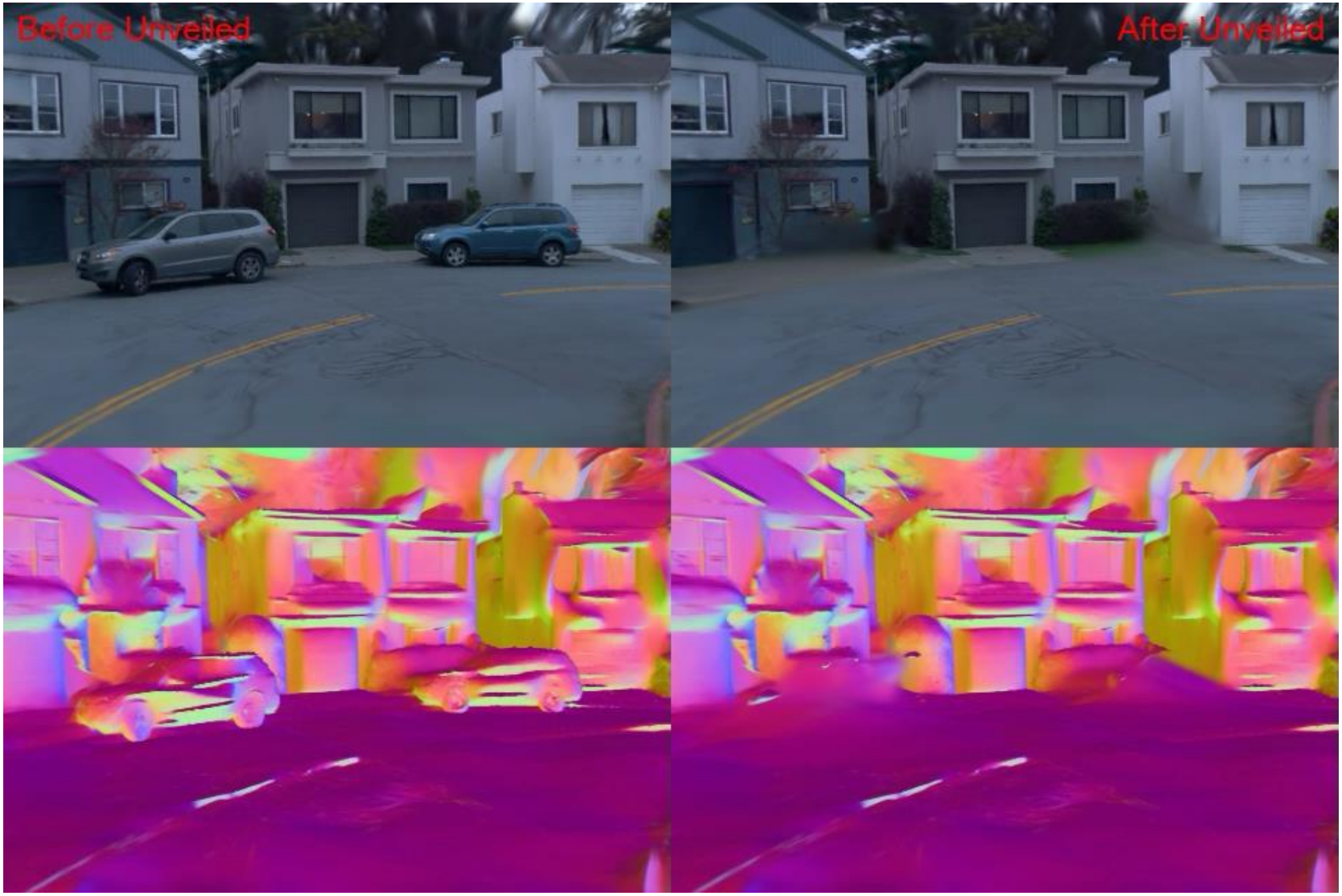
Remove the cars from the video captured by in-car cameras

Remove from both visual and geometry aspect

Time reversal inpainting



Inpaint T_n with T_{n+1} as reference

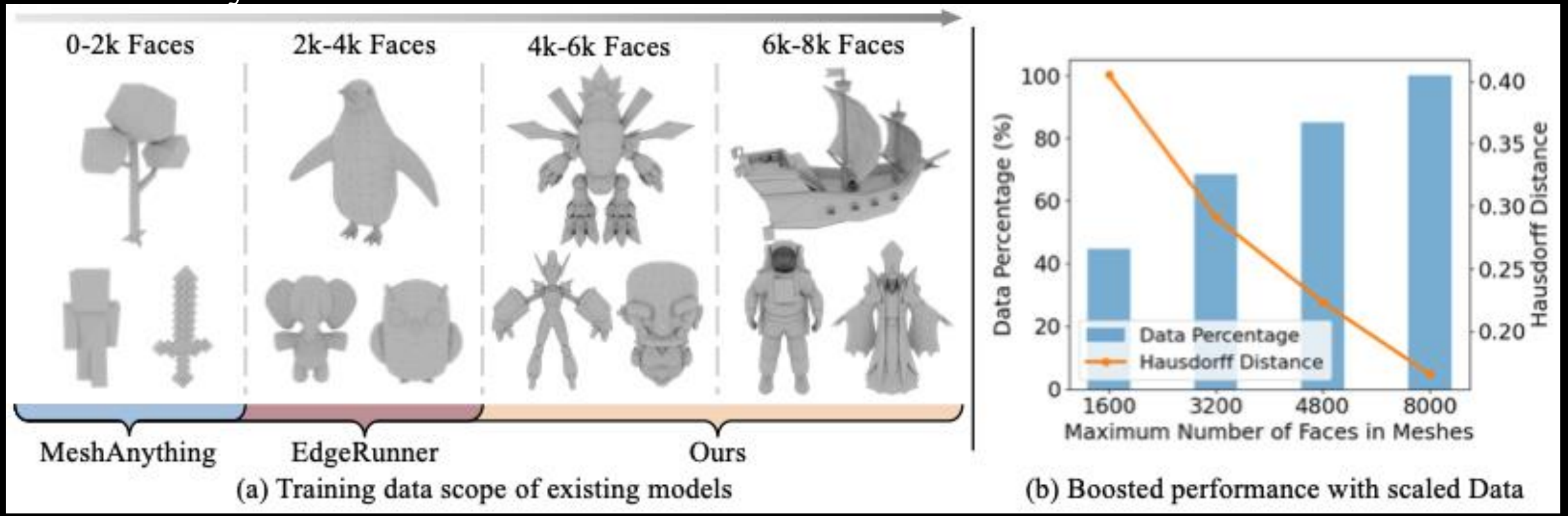

Before Unveiled
After Unveiled

Scaling Mesh Generation via Compressive Tokenization

CVPR 2025

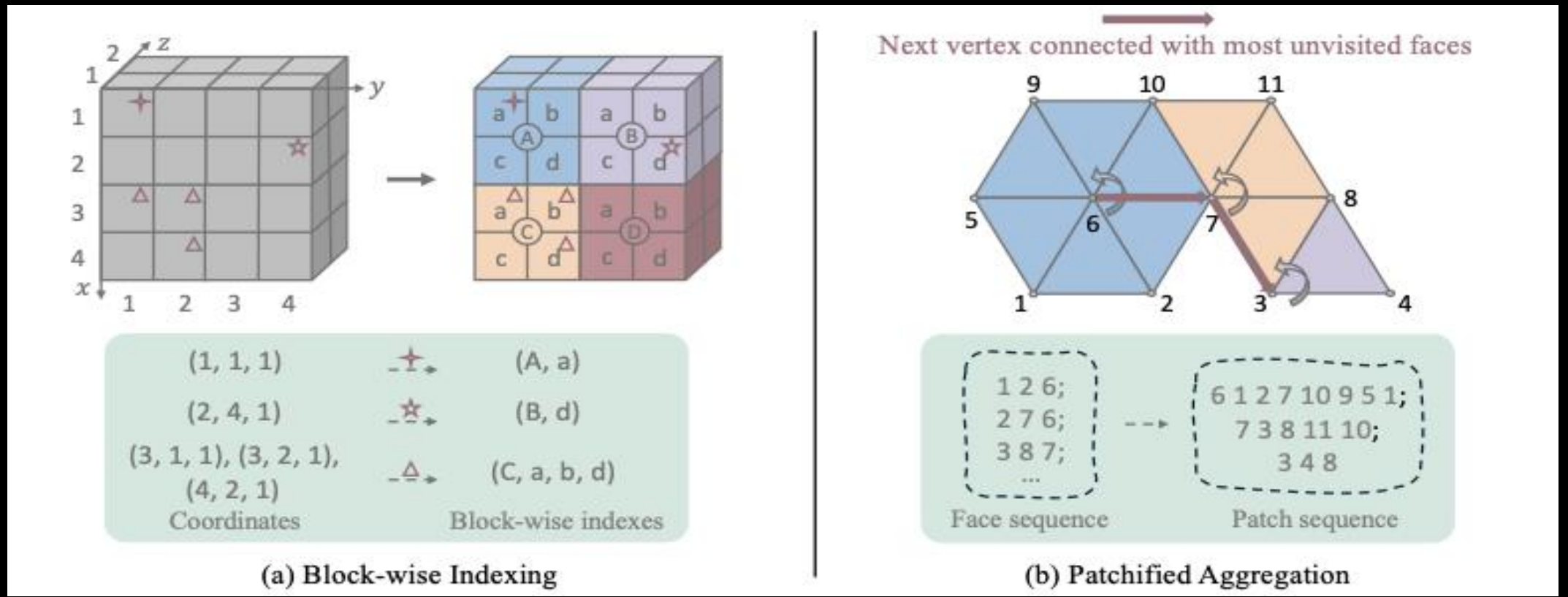
Scaling Mesh Generation via Blocked and Patchified Tokenization (BPT)

- Due to the inefficiency of the **mesh sequence tokenization**, the scale of the training data for the triangle mesh was limited, resulting in the generative model struggling to synthesize results with intricate details.



Scaling Mesh Generation via Blocked and Patchified Tokenization (BPT)

- **Vertices level**: Compared to the previous method of using three tokens to represent mesh vertices, we aim to use only one or two tokens to represent a mesh vertex
- **Face level**: We aim to design an order that minimizes the repetition of mesh vertices in the mesh sequence



Auto-Regressive Transformer Based Mesh Generation

- We use a standard auto-regressive Transformer with parameters θ to model the sequence with our tokenization and
- leverage cross-attention for conditioning. The token sequences are modeled with a standard auto-regressive Transformer with parameter θ, maximizing the log probability.

$$L(\theta) = \prod_{i=1}^{|P|} p(p_i | p_{1:i-1}, c; \theta),$$

- Point/image: Michelangelo encoder

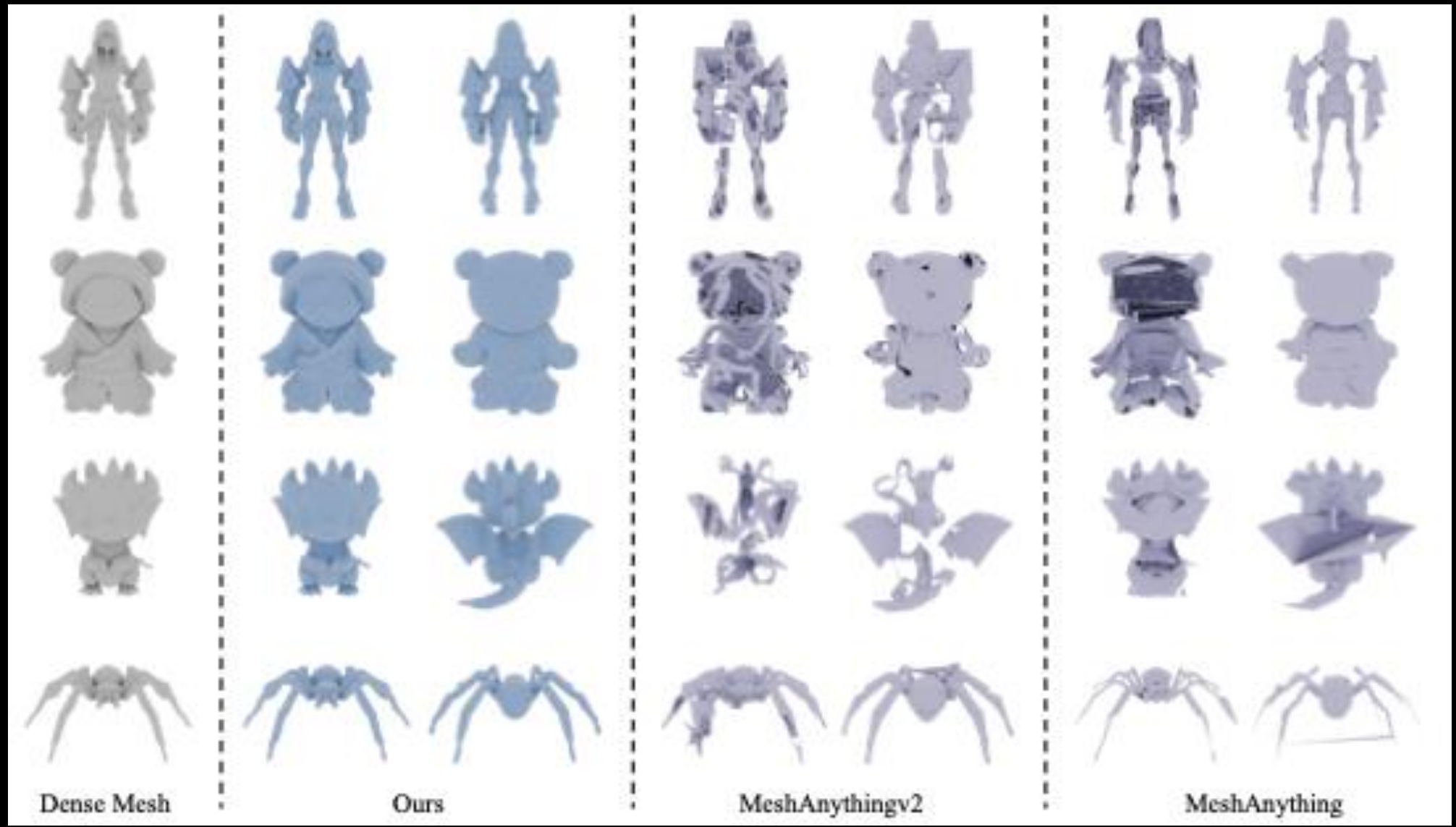
Evaluation

Table 2. **Quantitative comparison with baselines.** With the proposed BPT, our model can utilize meshes with many more faces, thus greatly improving the generation performance and robustness.

Method	Hausdorff Distance ↓	Chamfer Distance ↓
MeshAnything [4]	0.301	0.136
MeshAnythingv2 [5]	0.265	0.114
Ours	**0.166**	**0.094**

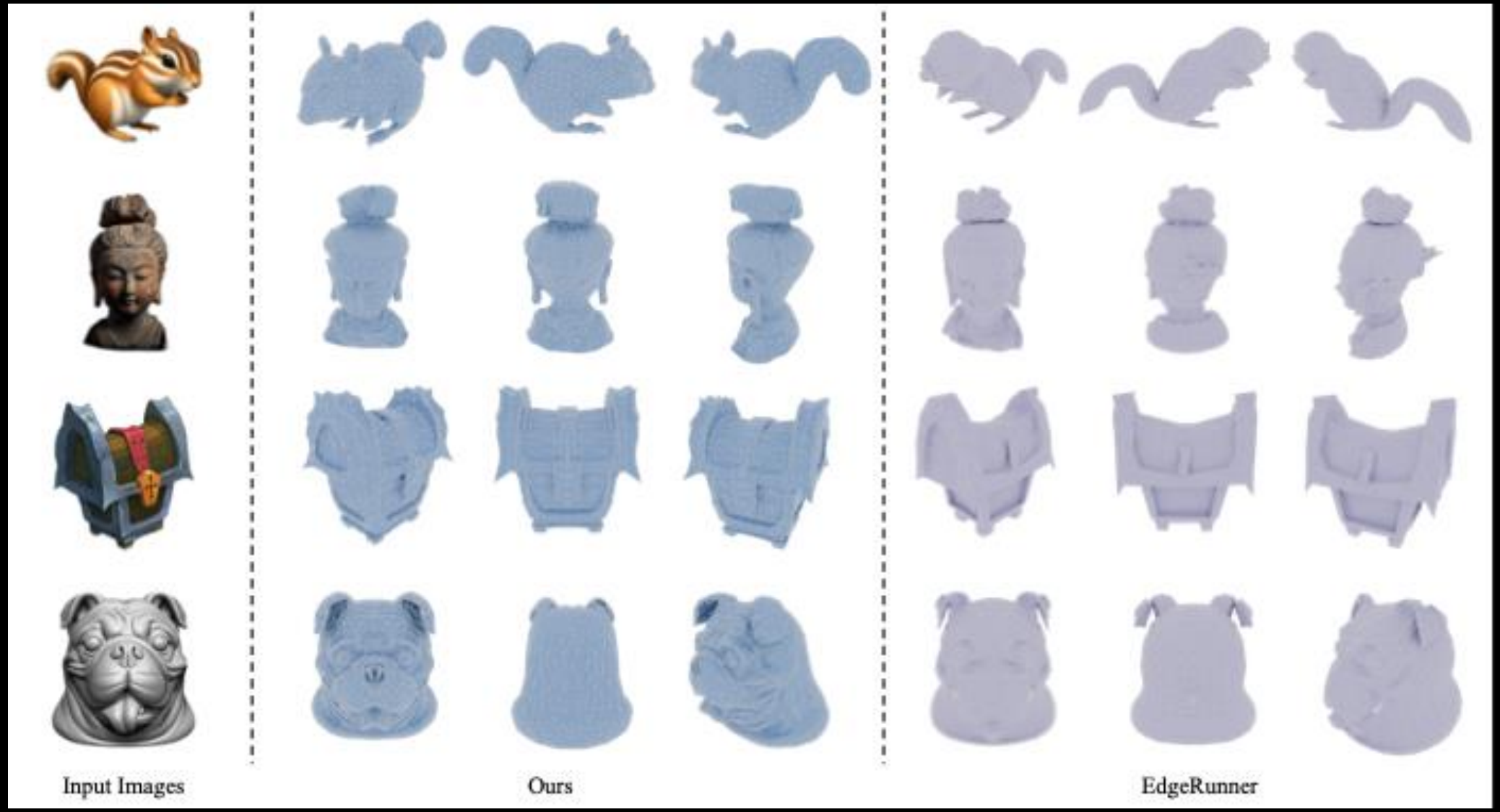
Scaling Mesh Generation via Compressive Tokenization

- Point Cloud Conditioned Generation



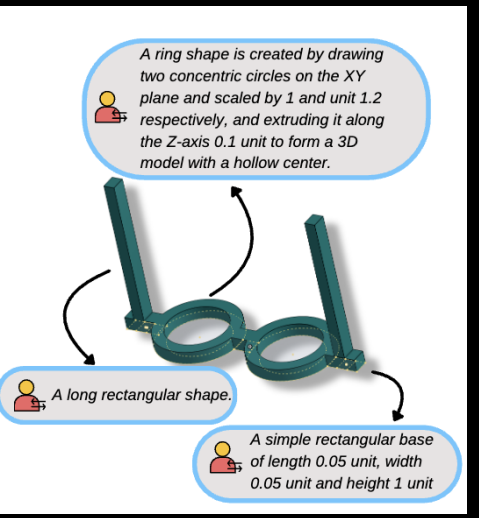
Scaling Mesh Generation via Compressive Tokenization

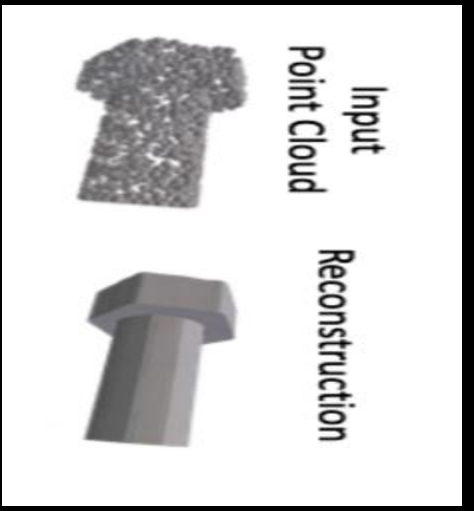
- Image Conditioned Generation



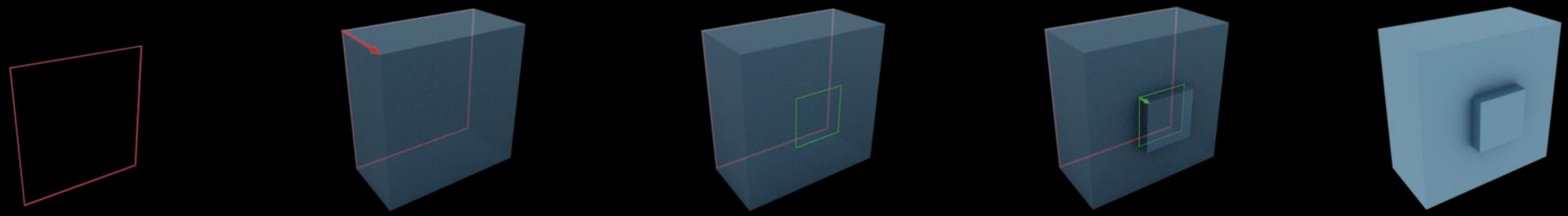
CAD-MLLM: Unifying Multimodality-Conditioned CAD Generation With MLLM

- Img2CAD and GenCAD have been proposed to generate a CAD model based on the input images.
- Text2CAD and Query2CAD have been proposed to generate a CAD model based on the text.
- Point2cyl and TransCAD have been proposed to generate a CAD model based on the point cloud.
- **Users' requirement: Design a solution that can work based on conditions of different modalities.**





Example of Command Sequence Representation



Sketch 1 → Extrude 1 → Sketch 2 → Extrude 2 → Union

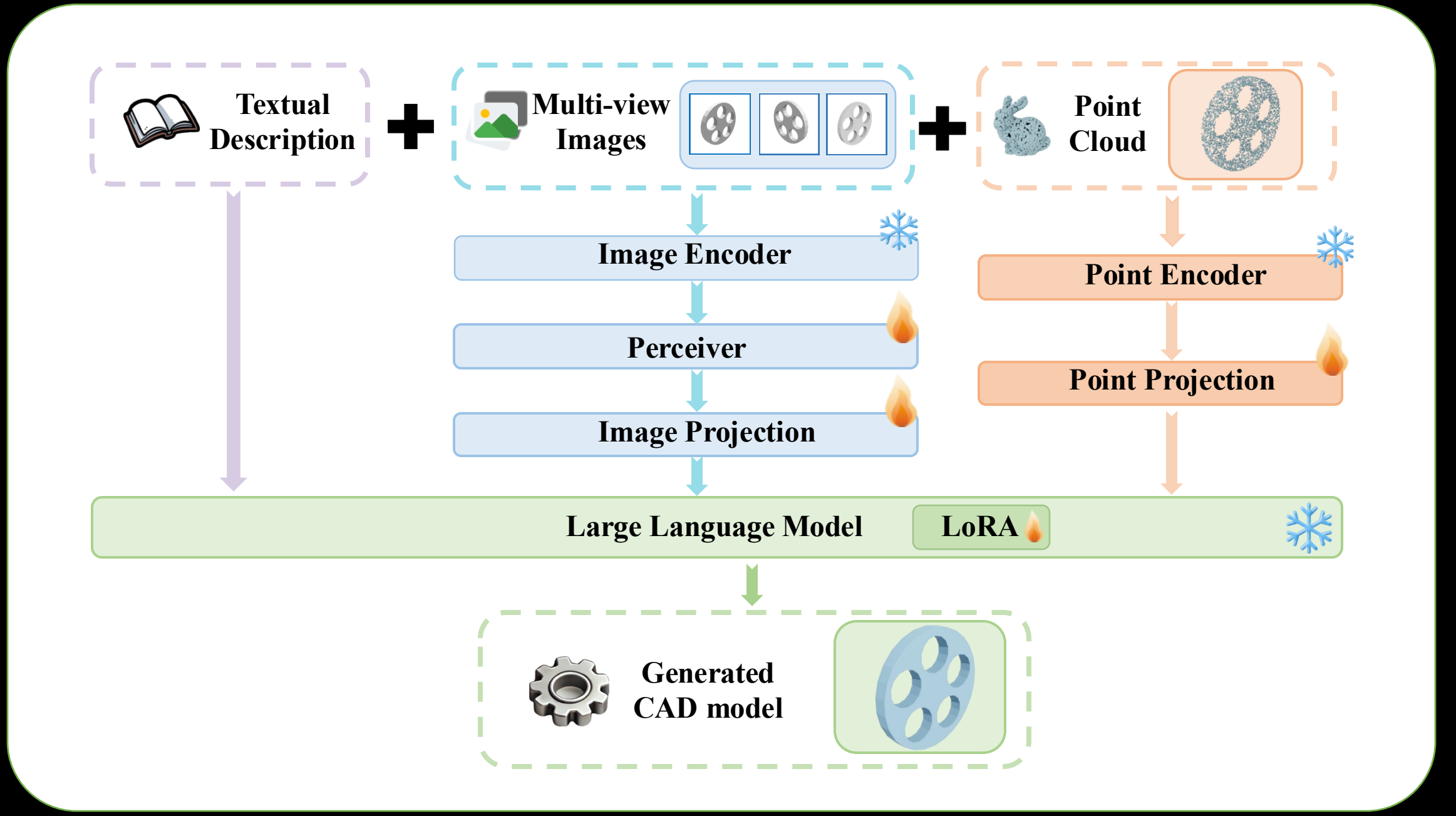
⟨SOS⟩ : Start of Sketch 1
⟨TOS⟩ : Line
⟨GOS⟩ : Position of Line
⟨EOC⟩ : End of Command

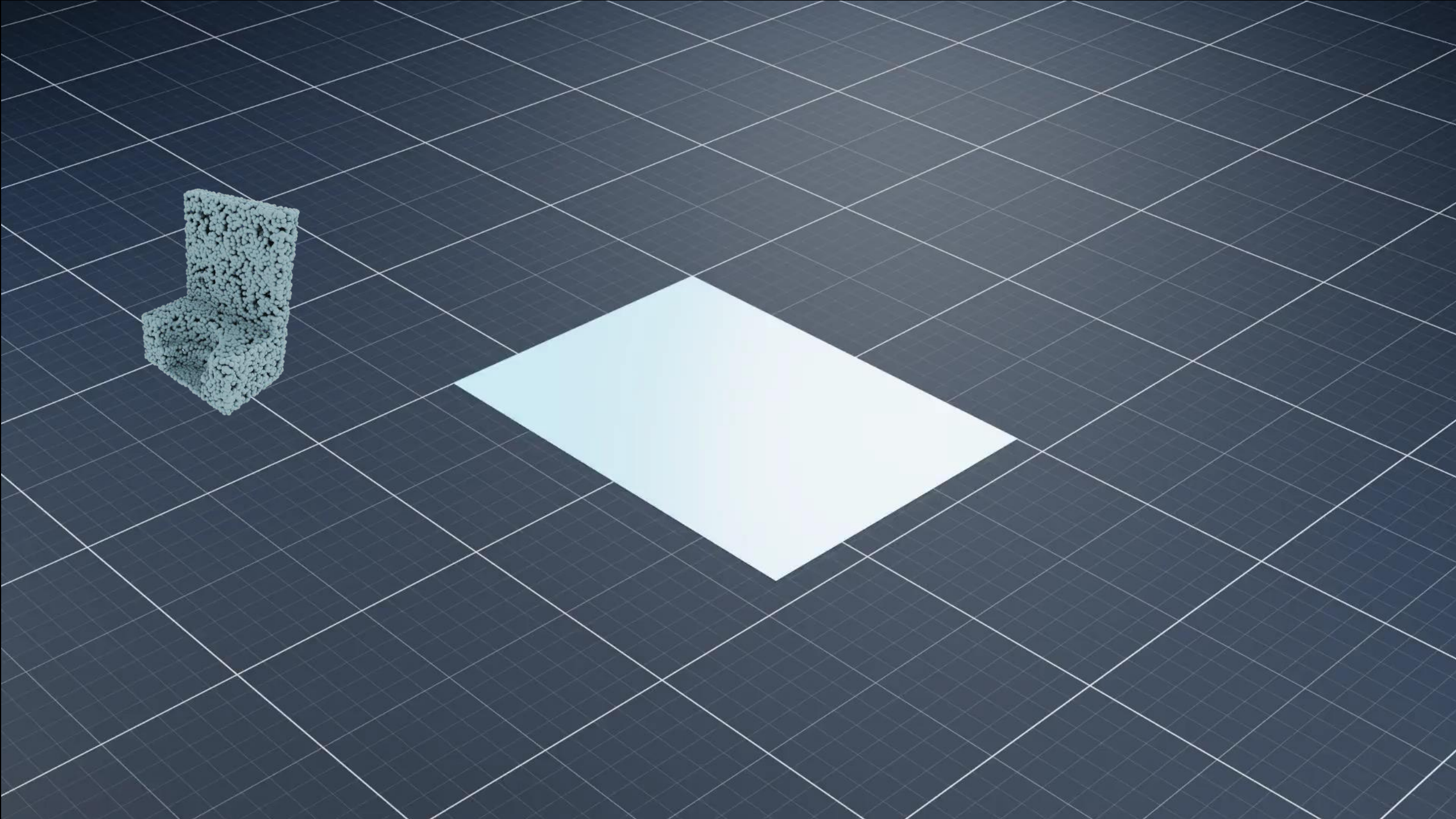
⟨SOE⟩ : Extrude Sketch 1
⟨GOE⟩ : Direction & Length
⟨KOE⟩ : Create body
⟨EOC⟩ : End of Command

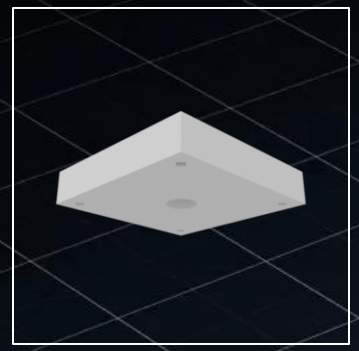
⟨SOS⟩ : Start of Sketch
⟨TOS⟩ : Line
⟨GOS⟩ : Position of Line
⟨EOC⟩ : End of Command

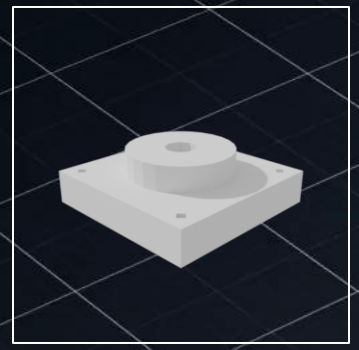
⟨SOE⟩ : Extrude Sketch 2
⟨GOE⟩ : Direction & Length
⟨KOE⟩ : Union
⟨EOC⟩ : End of Command

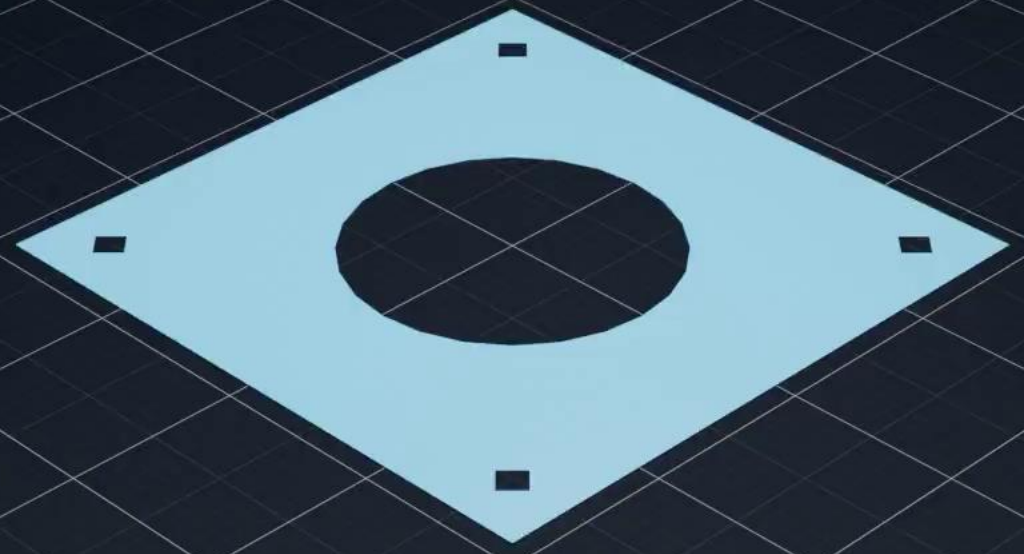
⟨EOS⟩ : End of Sequence

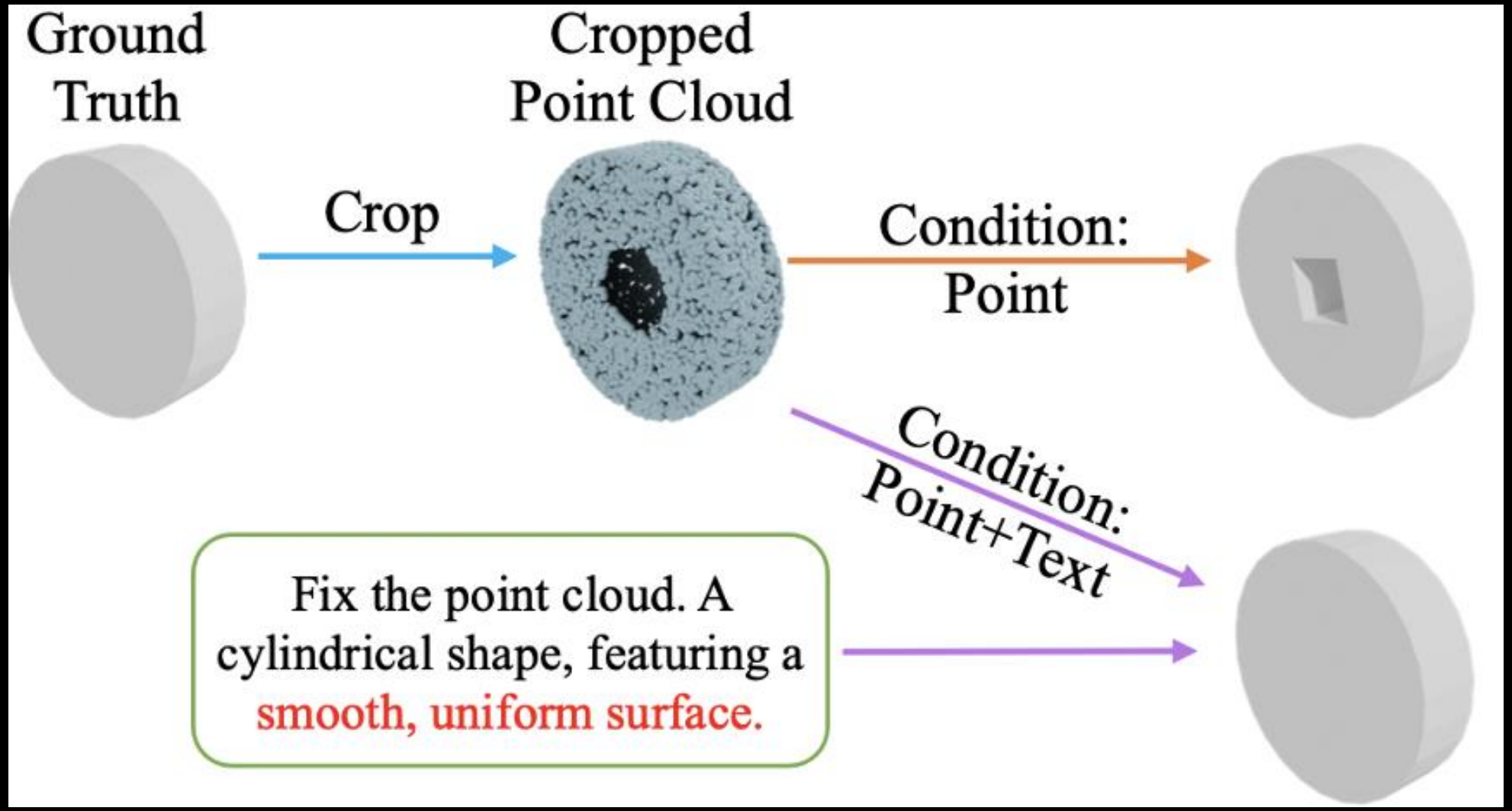

Textual Description
Multi-view Images
Point Cloud
Image Encoder
Perceiver
Image Projection
Point Encoder
Point Projection
Large Language Model
LoRA
Generated CAD model

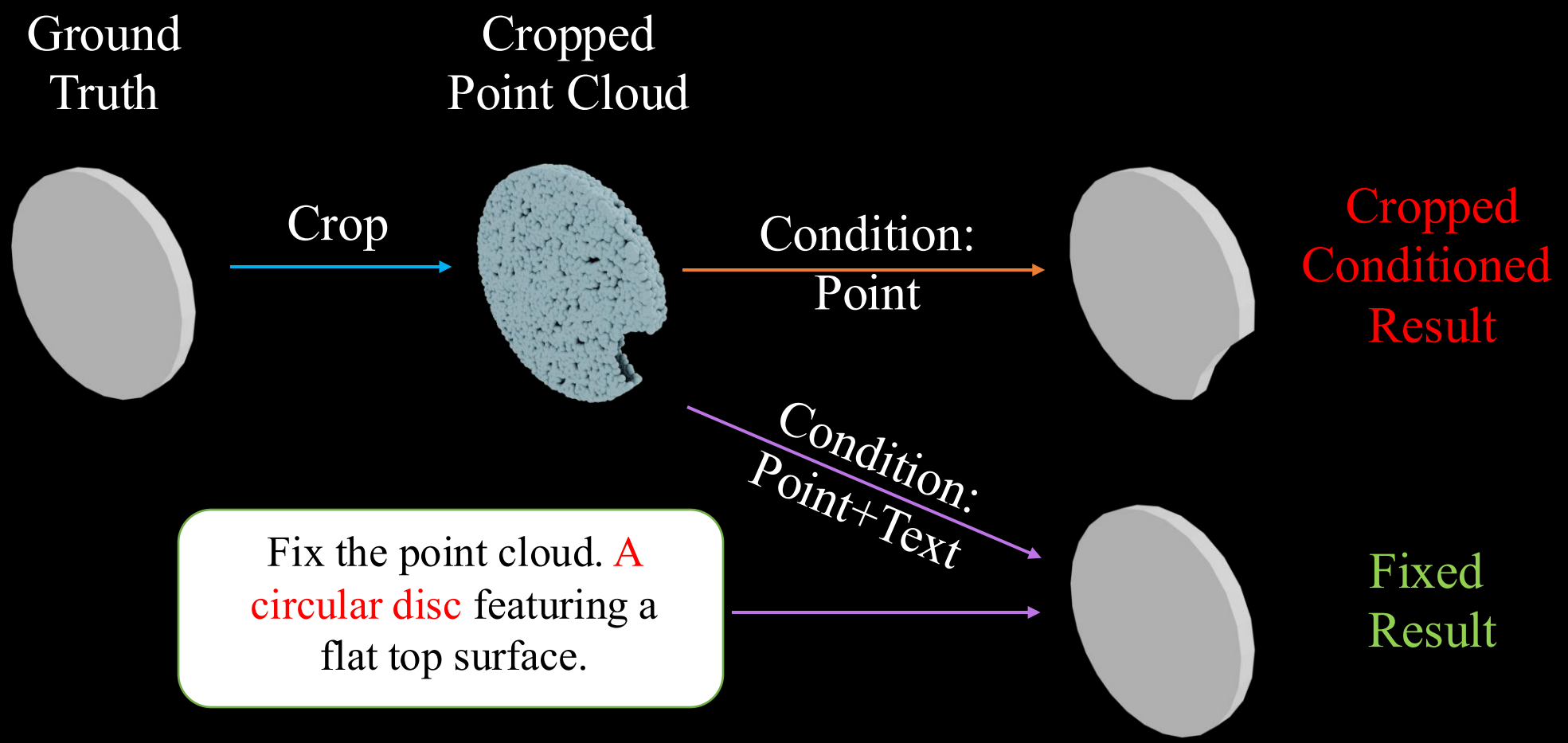









Ground Truth
Cropped Point Cloud
Crop
Condition: Point
Condition: Point+Text
Fix the point cloud. A cylindrical shape, featuring a smooth, uniform surface.



(a) Fix cropped point cloud with text

Summary

- Which representation is better for 3D ?
  - Application driven?
- How to leverage 3D geometric prior for reconstruction/generation?
- Learning from history.